Objects Before Words: Object-First Inductive Biases for Grounding Language in Child-View Video

Sathira Silva (sathira.silva@mbzuai.ac.ae)[1], Abrham Kahsay Gebreselasie[1], Muhammad Umer Sheikh[1], Kartik Kuckreja[1], Daniel Harari[2], & Muhammad Haris Khan[2]

[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

[2]Weizmann Institute of Science, Rehovot, Israel

Abstract

Learning grounded word meaning from natural experience requires resolving two ambiguities in infant-view recordings: *when* the named referent appears and *where* it is in a cluttered frame. In SAYCam-style data, caregiver speech is sparse and weakly synchronized with egocentric video, so single-frame contrastive pairing yields noisy positives in which the intended object is absent or entangled with distractors. We propose BabyMind, an object-first bias for child-view contrastive learning under sparse, noisy supervision. BabyMind extracts candidate object embeddings using an offline mask-based region interface, links candidates across a short utterance-centered window into lightweight object files via tracking, and aligns utterances to bags of object files with a prototype-space multiple-instance contrastive objective. Track-coherence and global-object agreement regularizers stabilize learning and transfer object-file structure into the global frame embedding used at evaluation. On SAYCam-S, BabyMind improves Labeled-S 15 forced-choice accuracy by +2.6 points over CVCL and yields consistent gains on in-vocabulary out-of-distribution benchmarks.

Keywords: grounded language learning; child-view video; egocentric vision; contrastive learning; multiple-instance learning; object files; prototype memory; SAYCam

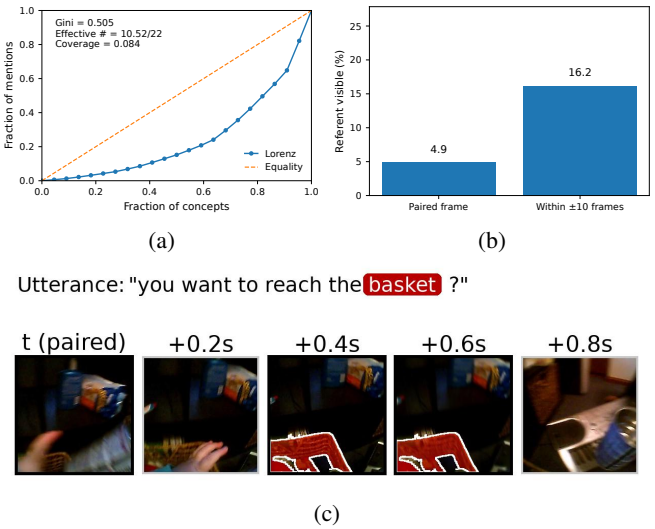

Figure 1: **SAYCam-S sparsity and misalignment.** (a) Long-tailed concept mentions. (b) Referents are often absent in the paired frame but present within a short window. (c) Example where the referent appears shortly after the paired time step.

Introduction

A central goal of grounded language learning is to explain how a learner can acquire word meanings from perceptual experience paired with sparse, weakly structured linguistic input (Harnad, 1990). This problem looks fundamentally different in early child learning than in curated image-caption corpora: the visual stream is egocentric, cluttered, partially occluded, and constantly moving, while caregiver speech is intermittent and only loosely synchronized with what is currently in view. In SAYCam-style data (Sullivan et al., 2021), this creates two recurring ambiguities: *when* the named referent appears and *where* it is in a crowded scene. These are not edge cases: concept mentions are long-tailed and imbalanced (Figure 1a), and a referent can be absent at the paired time step but present nearby in time (Figure 1b), as illustrated qualitatively in Figure 1c. These properties motivate an "objects before words" inductive bias. Classic accounts argue that robust recognition depends on intermediate perceptual structure that supports later interpretation (Mandler, 1992; Marr, 1982). Empirically, infants track object continuity under motion and occlusion and often generalize early nouns by shape (Baillargeon, 1987; Kellman & Spelke, 1983; Landau et al., 1988; Smith, 2003). By contrast, standard visual representations can over-rely on texture and background cues (Geirhos et al., 2019), which is especially problematic in egocentric scenes where context is pervasive and named referents can occupy only a small region. Child-View Contrastive Learning (CVCL) (Vong et al., 2024) takes an important step toward this setting by learning CLIP-style cross-modal embeddings (Radford et al., 2021) from utterances paired with child-view frames. However, CVCL trains on a single sampled frame per utterance, so positives can be noisy under temporal jitter, occlusion, and clutter. Whole-frame embeddings can also explain an utterance using background context or salient distractors, which can be especially harmful for rare concepts (Figure 1a).

Motivated by this perspective, we introduce BabyMind, an object-first inductive bias that augments CVCL with an auxiliary object-file pathway. BabyMind keeps the original CVCL global contrastive objective on the same single anchor frame per utterance and adds a short window of nearby frames used only for object-centric learning. From each frame, we extract *instance-level* candidates using an offline automatic mask generation (AMG) strategy, with a patch fallback to handle miss-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

ing or degenerate masks. We then link candidates across the window into short *object files* using lightweight tracking and align the utterance to the resulting *bag of tracked object files* with a multiple-instance contrastive objective (Maron & Lozano-Pérez, 1998). Crucially, this is not just relaxing a one-to-one pairing into a one-of-many relation: the latent alignment is constrained to tracked, instance-defined candidates and further stabilized by a small prototype memory plus auxiliary signals. Specifically, we compute MIL in prototype space (a learned codebook that encourages reusable appearance structure under long-tailed, noisy supervision), regularize tracks with a coherence loss across frames, and add a global-object agreement loss that transfers the selected object-file signal into the global frame embedding used at evaluation. Our contributions can be summarized as:

- BabyMind, an object-first extension of CVCL that addresses temporal and spatial ambiguity by aligning speech to tracked instance candidates within a short window, while keeping the original CVCL global objective and evaluation interface.
- Offline Automatic Mask Generation (AMG) for instance-level region masks and short-window object files via lightweight tracking for egocentric video.
- A prototype-space multiple-instance contrastive objective, with track coherence and global-object agreement to stabilize object selection and inject object-centric structure into the global embedding.
- Improved SAYCam-S grounding on Labeled-S 15 and consistent (if modest) gains on IV out-of-distribution evaluations under the CVCL protocol.

Related Work

Grounded language learning from child-view video. Long-form egocentric corpora such as SAYCam provide a naturalistic testbed for grounded learning under clutter, occlusion, and weak temporal coupling between caregiver speech and the child's visual stream (Sullivan et al., 2021). CVCL demonstrates that CLIP-style vision-language alignment can be learned in this regime using utterance-frame pairing and contrastive objectives (Radford et al., 2021; Vong et al., 2024), complementing evidence that high-level visual representations can emerge from child-like inputs even with limited built-in structure (Orhan & Lake, 2024; Orhan et al., 2020). A central challenge in this setting is that supervision is intrinsically ambiguous: the relevant object may fall outside the sampled frame or be visually confounded by background context. Complementary benchmark evidence (Chen et al., 2026) suggests that perceptual competence remains a bottleneck for current multimodal systems even on tasks that do not primarily depend on language, motivating stronger perceptual organization for grounding.

Object-centric structure and region interfaces in vision-language. A large body of vision-language work injects object-level structure via region proposals or detector features, which has become a standard interface for captioning/VQA and later for region-aware pretraining (Anderson et al., 2018; Li et al., 2020; Tan & Bansal, 2019; Zhang et al., 2021). Related ideas appear in self-supervised learning, where region/mask structure is used to reduce shortcut reliance and to define localized learning targets (Hénaff et al., 2021). In parallel, object-centric representation learning has explored architectural constraints that decompose scenes into entities without labels (Burgess et al., 2019; Locatello et al., 2020). Recent foundation segmentation models make it feasible to obtain instance masks as a general-purpose perceptual prior across domains (Kirillov et al., 2023), aligning with results showing that perceptual inductive biases can materially shape what contrastive learning captures (Zhao et al., 2025).

Ambiguous instance selection, prototype memories, and temporal persistence. When supervision applies to a set of candidates rather than a single labeled instance, multiple-instance learning (MIL) provides a principled framework for latent instance selection (Maron & Lozano-Pérez, 1998). Prototype- and clustering-based mechanisms are also widely used to stabilize learning and encourage reusable structure in self-supervision, including memory-bank approaches and online assignment/clustering methods (Caron et al., 2018, 2020, 2021; van den Oord et al., 2017; Wu et al., 2018). Finally, learning from temporal continuity is a longstanding theme in unsupervised learning from video (Sermanet et al., 2017; Wiskott & Sejnowski, 2002), and cognitive accounts formalize persistence via token-like "object files" maintained across time (Kahneman et al., 1992). These threads motivate combining set-based alignment with reusable prototype structure and short-range temporal persistence when learning grounded meaning from egocentric video.

Methodology

Overview. We learn aligned representations of child-view video and caregiver speech in the Child-View Contrastive Learning (CVCL) setting (Vong et al., 2024) on SAYCam (Sullivan et al., 2021). In egocentric video, an utterance may refer to an object whose time of appearance and pixels are unknown, so single-frame contrastive pairing is sensitive to temporal mismatch and background clutter. BabyMind resolves this ambiguity by introducing an object-file pathway inspired by token-level persistence (Kahneman et al., 1992): (i) a fixed region interface from offline automatic mask generation (AMG) (Kirillov et al., 2023), (ii) short-window object files formed by greedy cross-frame linking, and (iii) a prototype-space multiple-instance contrastive objective (Maron & Lozano-Pérez, 1998). Two regularizers shape this pathway: track coherence in prototype space (temporal stability (Wiskott & Sejnowski, 2002)) and global-object agreement that transfers object-file structure into the global embedding used at evaluation.

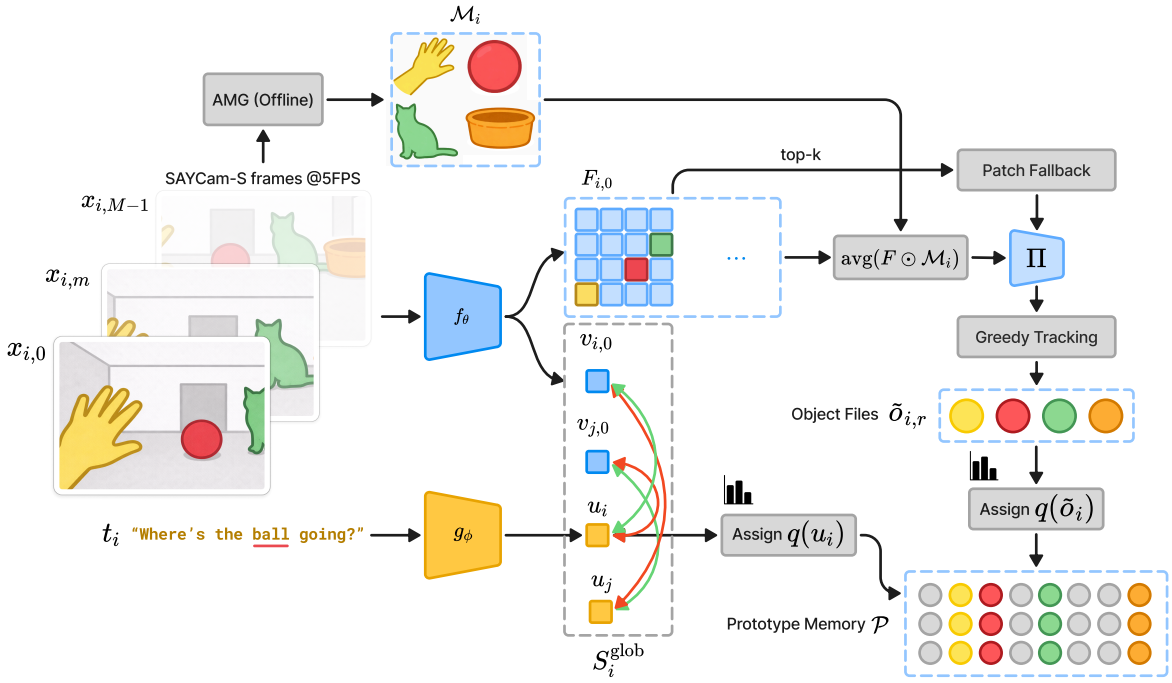

Figure 2: **Method overview.** Each example contains an utterance t_i and a window of M utterance-aligned frames $\{x_{i,m}\}_{m=0}^{M-1}$, with $m = 0$ the anchor used by the global CVCL objective. We extract mask-defined region embeddings from feature maps, merge them into object files across the window, align text to the resulting bag via prototype-space MIL, regularize tracks via prototype consistency across frames, and align the anchor global embedding to a prototype reconstruction of the selected object file.

Problem setup

A minibatch contains B examples indexed by $i \in \{1, \ldots, B\}$. Example i contains an utterance transcript t_i and a temporal window of M frames $\{x_{i,m}\}_{m=0}^{M-1}$. The anchor frame $x_{i,0}$ is sampled uniformly at random from the utterance-aligned frame set as in CVCL and is used by $\mathcal{L}_{\text{glob}}$. The remaining $M-1$ frames are additional samples from the same utterance-aligned set and are used only for object-file learning. A text encoder g_ϕ and a vision encoder f_θ produce normalized embeddings and a spatial feature map:

$$u_i = \frac{g_\phi(t_i)}{\|g_\phi(t_i)\|_2} \in \mathbb{S}^{d-1}, \tag{1}$$

$$v_{i,m} = \frac{f_\theta^{\text{glob}}(x_{i,m})}{\|f_\theta^{\text{glob}}(x_{i,m})\|_2} \in \mathbb{S}^{d-1}, \tag{2}$$

$$F_{i,m} = f_\theta^{\text{map}}(x_{i,m}) \in \mathbb{R}^{C \times H \times W}. \tag{3}$$

We denote $v_i = v_{i,0}$ and use cosine similarity $s(a, b) = a^\top b$ for normalized vectors.

Global CVCL objective

CVCL aligns utterances to anchor frames with a symmetric contrastive loss (Vong et al., 2024). Define

$$S_{ij}^{\text{glob}} = \frac{s(u_i, v_j)}{\tau_{\text{glob}}}, \tag{4}$$

and optimize

$$\mathcal{L}_{\text{glob}} = \frac{1}{2B} \sum_{i=1}^{B} \Big(\text{CE}(S_{i,:}^{\text{glob}}, i) + \text{CE}((S^{\text{glob}})_{i,:}^\top, i) \Big), \tag{5}$$

with temperature $\tau_{\text{glob}} > 0$.

Region candidates and object embeddings

For each frame $x_{i,m}$ we use a set of binary masks $\mathcal{M}_{i,m} = \{M_{i,m,r}\}_{r=1}^{R_{i,m}}$ and downsample them to feature-map resolution, providing $\tilde{M}_{i,m,r} \in [0,1]^{H \times W}$. Let $F_{i,m}^{p,q} \triangleq F_{i,m}(:, p, q) \in \mathbb{R}^C$ denote the feature vector at location (p, q).

Masked average pooling. We compute a region descriptor by masked averaging:

$$f_{i,m,r} = \frac{\sum_{p,q} \tilde{M}_{i,m,r}[p,q]\, F_{i,m}^{p,q}}{\sum_{p,q} \tilde{M}_{i,m,r}[p,q] + \varepsilon} \in \mathbb{R}^C, \tag{6}$$

with $\varepsilon > 0$ for numerical stability. A projection head $\Pi : \mathbb{R}^C \to \mathbb{R}^d$ maps region descriptors to the shared space:

$$o_{i,m,r} = \frac{\Pi(f_{i,m,r})}{\|\Pi(f_{i,m,r})\|_2} \in \mathbb{S}^{d-1}. \tag{7}$$

Patch candidates for missing or degenerate masks. If a mask becomes empty after downsampling, we replace it with

the feature at its centroid location (in feature-map coordinates). If a frame has no surviving masks, we select K_{patch} salient spatial locations from $F_{i,m}$ and treat their projected features as patch candidates. This guarantees every frame provides a valid candidate set.

Object files via short-window tracking

We merge per-frame candidates into short object files using greedy similarity tracking. Within each example i, candidates are assigned to the most similar existing track if cosine similarity exceeds a threshold; otherwise a new track is created. Each track embedding is the ℓ_2-normalized mean of its assigned candidate embeddings. We keep at most $R_{\max}$ tracks and denote track embeddings by $\tilde{O}_i = \{\tilde{o}_{i,1}, \ldots, \tilde{o}_{i,R_i}\}$.

Null track and padding. We append a learnable null embedding $\tilde{o}_\varnothing \in \mathbb{S}^{d-1}$ to represent "no grounded referent in the window". For batching, each bag is padded to size $R_{\max} + 1$ with a validity mask $m_{i,r} \in \{0, 1\}$ over padded entries (the null entry is always valid).

Prototype memory

We maintain a prototype codebook $\mathcal{P} = \{p_k\}_{k=1}^K \subset \mathbb{S}^{d-1}$. Given $x \in \mathbb{S}^{d-1}$, the soft prototype assignment is:

$$q(x)_k = \text{softmax}_k\left(\frac{s(x, p_k)}{\tau_{\text{proto}}}\right), \qquad k \in \{1, \ldots, K\}, \tag{8}$$

with temperature $\tau_{\text{proto}} > 0$. We use $q_i^{\text{txt}} = q(u_i)$ and $q_{i,r}^{\text{trk}} = q(\tilde{o}_{i,r})$. Prototypes are updated online by EMA using Sinkhorn-normalized assignments as in SwAV (Caron et al., 2020).

Prototype-space MIL alignment

We align each utterance to the bag of object files via MIL (Maron & Lozano-Pérez, 1998). Define instance similarity

$$a_{ijr} = (q_i^{\text{txt}})^\top q_{j,r}^{\text{trk}}, \tag{9}$$

and aggregate over tracks with masked log-sum-exp pooling:

$$S_{ij}^{\text{mil}} = \tau_{\text{MIL}} \log \sum_{r=1}^{R_{\max}+1} m_{j,r} \exp\left(\frac{a_{ijr}}{\tau_{\text{MIL}}}\right), \tag{10}$$

where $\tau_{\text{MIL}} > 0$ controls pooling hardness. We apply a symmetric in-batch contrastive loss:

$$\mathcal{L}_{\text{mil}} = \frac{1}{2B} \sum_{i=1}^{B} \Big(\text{CE}(S_{i,:}^{\text{mil}}, i) + \text{CE}((S^{\text{mil}})_{i,:}^\top, i)\Big). \tag{11}$$

Regularizers

Track coherence. Object files are intended to represent the same underlying entity across frames. For each example i, let $\mathcal{T}_i$ be its set of non-null tracks. For each track $r \in \mathcal{T}_i$, let $\mathcal{A}_{i,r} \subseteq \{0, \ldots, M-1\}$ be the set of frames in which the track has an assigned per-frame candidate (from the tracking step), and let $q_{i,m,r}$ be the prototype assignment of that per-frame candidate. We regularize the per-frame assignments to match the track assignment using KL divergence with a stop-gradient teacher:

$$\begin{aligned}
\ell_{i,r}^{\text{coh}} &= \frac{1}{|\mathcal{A}_{i,r}|} \sum_{m \in \mathcal{A}_{i,r}} \text{KL}\Big(\text{sg}[q_{i,r}^{\text{trk}}] \,\|\, q_{i,m,r}\Big), \\
\ell_i^{\text{coh}} &= \frac{1}{|\mathcal{T}_i|} \sum_{r \in \mathcal{T}_i} \ell_{i,r}^{\text{coh}}, \\
\mathcal{L}_{\text{coh}} &= \frac{1}{B} \sum_{i=1}^{B} \ell_i^{\text{coh}},
\end{aligned} \tag{12}$$

where $\text{sg}[\cdot]$ denotes stop-gradient.

Global-object agreement. CVCL-style evaluation uses the global embedding v_i. To transfer object-file structure into this representation, we select the best non-null track

$$r_i^\star = \arg\max_{r \in \mathcal{T}_i} (q_i^{\text{txt}})^\top q_{i,r}^{\text{trk}}, \tag{13}$$

reconstruct a continuous embedding from its prototype mixture,

$$\hat{o}_i = \frac{\sum_{k=1}^K q(\tilde{o}_{i,r_i^\star})_k \; p_k}{\left\| \sum_{k=1}^K q(\tilde{o}_{i,r_i^\star})_k \; p_k \right\|_2}, \tag{14}$$

and minimize cosine distance:

$$\mathcal{L}_{\text{go}} = \frac{1}{B} \sum_{i=1}^{B} \big(1 - s(v_i, \hat{o}_i)\big). \tag{15}$$

Training objective

We optimize (θ, ϕ), the projection head Π, and the null embedding. The prototype memory is updated online as described above. The full objective is:

$$\mathcal{L} = \lambda_{\text{glob}} \mathcal{L}_{\text{glob}} + \lambda_{\text{MIL}} \mathcal{L}_{\text{mil}} + \lambda_{\text{coh}} \mathcal{L}_{\text{coh}} + \lambda_{\text{go}} \mathcal{L}_{\text{go}}, \tag{16}$$

with nonnegative weights $\lambda_\cdot$. Unless otherwise stated, evaluation uses the global embedding v (the CVCL interface).

Experiments and Analyses

We evaluate BabyMind in the Child-View Contrastive Learning (CVCL) setting (Vong et al., 2024): self-supervised learning from egocentric developmental video paired with caregiver speech (SAYCam-S (Sullivan et al., 2021)), followed by CVCL-style 4-way forced-choice tests of vision-language alignment. Our experiments ask: (i) does object-file supervision improve SAYCam grounding? (ii) does it generalize under CVCL's in-vocabulary (IV) OOD protocol? and (iii) what diagnostic structure emerges in the learned object/prototype representations?

Evaluation (CVCL forced-choice). Each trial contains a target word and four candidate images (one target, three foils). The model selects the image whose embedding has the highest cosine similarity to the text embedding; we report average accuracy (%), with chance at 25%.

Method	Ball	Basket	Car	Cat	Chair	Computer	Crib	Door	Foot	Hand	Paper	Puzzle	Stairs	Table	Window
CVCL	**86**	29	94	**58**	66	94	94	74	26	23	78	**91**	**87**	**80**	73
BabyMind	75	**42**	**95**	56	**70**	94	**98**	**78**	**39**	**31**	**86**	87	85	76	**80**

Table 1: **Per-category accuracy on Labeled-S 15.** 4-way forced-choice accuracy (%).

Method	Avg (%)	Δ
CVCL (reprod)	70.20	-
BabyMind	**72.80**	**+2.60**

Table 2: **Labeled-S 15 (SAYCam-S).** 4-way forced-choice average accuracy. Δ is relative to CVCL.

Variant	Avg (%)
CVCL	70.20
BabyMind (full)	**72.80**
w/o prototype MIL (object-centric branch)	70.30
w/o object files (no tracking)	72.10
w/o track coherence (λ_{coh}=0)	72.50
w/o global-object agreement (λ_{go}=0)	71.20

Table 3: **Ablations on Labeled-S 15.** 4-way forced-choice average accuracy.

Training and evaluation setup

Implementation details. All models are implemented in PyTorch using PyTorch Lightning and trained with DistributedDataParallel on 4 AMD MI210 GPUs (batch size 8 per GPU; global batch size 32). For both the global CVCL loss and the prototype-space MIL loss, we gather all embeddings across GPUs (along with the MIL validity masks) to utilize a shared in-batch negative set. Precomputed AMG masks are loaded from a cached prepacked format, and we apply the same random geometric augmentations jointly to frames and masks via an image-mask transform (non-geometric appearance augmentations are applied to images only). We reproduce the CVCL baseline within the same codebase and distributed setup for a controlled comparison. Unless otherwise stated, for all the experiments (including CVCL reproduced results), we use AdamW with a learning rate 4×10^{-4} and weight decay 0.1, together with global temperature τ_{glob} = 0.07 as in CVCL, and report seed-0 results from the checkpoint with the best validation loss.

Utterance-frame pairing. We follow CVCL's preprocessing and pairing procedure (Vong et al., 2024): frames are sampled at 5 FPS, and each utterance is paired with all frames until the next utterance (capped at 32 frames). For the global CVCL objective, we sample a single *anchor* frame uniformly at random from this utterance-aligned set, exactly as in CVCL. BabyMind uses the same anchor frame for $\mathcal{L}_{\text{glob}}$ and additionally samples $M-1$ extra frames from the same utterance-aligned set to form a M-frame window used only by the object-file pathway. For tracking, we process the window frames in chronological order.

Generalization under CVCL IV/OOV protocol

Object candidates. We precompute SAM-style automatic mask generation (AMG) proposals offline. For each frame, we load up to 24 masks, downsample them to the backbone feature-map resolution, and filter tiny masks (area < 1% of the feature-map grid). If a frame has no valid masks after filtering, we construct patch candidates from the same feature map (top-K_{patch}=4 locations; patch radius 1 in feature map coordinates), ensuring that every frame yields a valid candidate set.

BabyMind hyperparameters. Prototype memory: K=64, τ_{proto}=0.07, EMA decay 0.99 with Sinkhorn balancing (3 iterations, ϵ=0.05). Object files: greedy tracking threshold 0.55, maximum of 16 tracks per window. Loss weights: λ_{glob}=1.0. λ_{MIL}=0.10, λ_{coh}=0.05, λ_{go}=0.05, with τ_{MIL}=0.05.

SAYCam grounding on Labeled-S 15

Our primary SAYCam evaluation is CVCL's manually filtered **Labeled-S 15** benchmark (Vong et al., 2024) (15 mutually exclusive categories; 100 forced-choice trials per category, 1500 total). BabyMind improves average accuracy by +2.60 points over CVCL (Table 2). Gains are concentrated in categories where the referent is often small, partially visible, or embedded in clutter (e.g., `Basket`, `Foot`, `Hand`, `Paper`, `Window`; Table 1), consistent with object-file MIL providing a cleaner supervisory signal than whole-frame alignment. We observe decreases on `Ball` (and modest drops on `Puzzle`/`Table`), suggesting that when global appearance cues are already strong, imperfect masks/tracks can introduce competing gradients.

Ablations on SAYCam-S

We ablate BabyMind components on Labeled-S 15 by removing each mechanism while keeping the rest fixed (Table 3). Two results stand out. First, removing prototype-space MIL (the object-centric branch) reduces performance to the CVCL baseline, showing that improvements are driven by object-file supervision rather than incidental training differences. Second, global-object agreement is the most important auxiliary term for downstream performance: disabling it costs 1.6 points, consistent with its role in transferring object-file structure into the global embedding used by forced-choice evaluation. Tracking into object files provides a smaller but consistent gain (0.7 points), and track coherence contributes mild regularization (0.3 points).

Dataset	Method	#IV	#OOV	OOD Avg (%)	OOV Avg (%)	Total Avg (%)
Konkle Object Categories	CVCL	65	135	34.32	**53.11**	47.00
	BabyMind	65	135	**35.41**	52.87	**47.20**
COCO (categories)	CVCL	70	10	32.78	**42.50**	34.00
	BabyMind	70	10	**33.14**	42.36	**34.29**

Table 4: **OOD (IV) and OOV generalization.** IV/OOD results use the CVCL 4-way forced-choice protocol with global embeddings. #IV/#OOV are label counts after the vocabulary filter; Total Avg is label-count-weighted over IV and OOV.

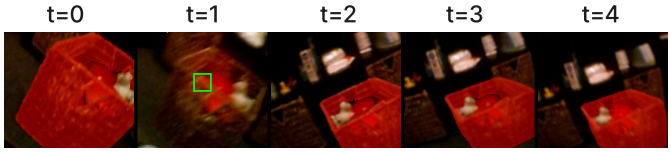

Figure 3: **Object-file robustness under mask dropouts.** A 5-frame window (M=5) for one tracked object file. Red overlays indicate selected AMG mask candidates when available. At t=1, no valid mask remains after filtering; the model falls back to a patch candidate (green box), preserving temporal coherence of the track.

We evaluate beyond SAYCam-S on Konkle Object Categories and COCO categories (Lin et al., 2014) using CVCL's in-vocabulary (IV) vs. out-of-vocabulary (OOV) split. IV labels use the same 4-way forced-choice test as on SAYCam-S (global embeddings), while OOV labels use a CLIP Dissect-style unit probing protocol (Oikarinen & Weng, 2022). For Konkle, we cap at 200 images per label and use 5 repeats per image (resampling foils). For COCO, we cap at 200 instances per label and evaluate instance crops, filtering by minimum box area (32×32) and minimum side length (16 px). BabyMind yields small but consistent IV/OOD gains on both datasets with essentially unchanged OOV probing (Table 4), matching its design: object-file supervision primarily strengthens the global embedding used by forced-choice recognition.

Qualitative and diagnostic analyses

Tracking through mask dropouts. Figure 3 illustrates one tracked object file across a 5-frame window. When AMG masks are unavailable after filtering (here at t=1), BabyMind falls back to a patch candidate (green box), maintaining a coherent track and keeping the object-file pathway well-defined under common egocentric failures (motion blur, occlusion, off-center framing).

Prototype memory diagnostics. Figure 4 summarizes prototype behavior: the usage distribution (effective #K and Gini) and example top-activating tracked object files for several prototypes. Prototype usage remains well-spread (no collapse), and the retrieved tracks show distinct recurring appearance/region patterns rather than a single dominant code.

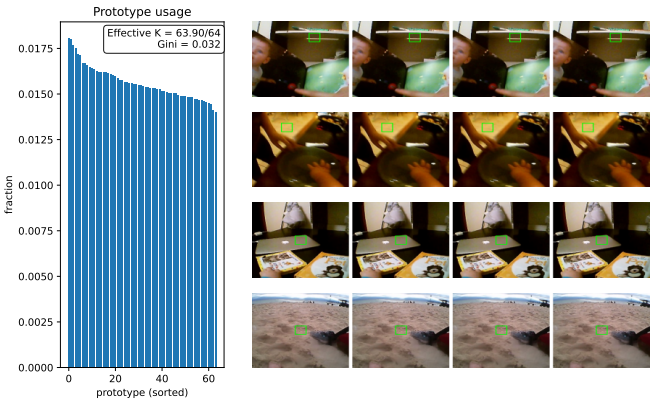

Figure 4: **Prototype diagnostics.** Left: prototype usage histogram (sorted), with effective #prototypes and Gini coefficient. Right: for several frequently used prototypes, the top-activating tracked object file mined from training data, shown as a 4-frame strip.

Conclusion

We introduced BabyMind, an object-first inductive bias for learning grounded word meaning from child-view video paired with sparse caregiver speech. Rather than treating utterance-frame alignment as a single-frame problem, BabyMind uses an offline region interface, merges candidates across a short utterance-centered window into lightweight *object files*, and aligns language to these latent candidates via a prototype-space multiple-instance contrastive objective. Two simple regularizers, track coherence and global-object agreement, help stabilize the object pathway and transfer its signal into the global embedding used by CVCL-style evaluation. Empirically, BabyMind improves SAYCam-S grounding on Labeled-S 15 and yields consistent (if modest) gains on in-vocabulary OOD benchmarks under the CVCL protocol, while leaving OOV unit-probing performance largely unchanged. Overall, the results support an "objects before words" perspective: even coarse, short-window perceptual organization can reduce temporal and spatial ambiguity in egocentric data and provide cleaner targets for early word grounding. Future work should reduce reliance on offline masks, extend object persistence beyond short windows, and evaluate robustness across additional children, environments, and more complex linguistic contexts.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Baillargeon, R. (1987). Object permanence in $3\frac{1}{2}$- and $4\frac{1}{2}$-month-old infants. *Developmental Psychology*, *23*(5), 655–664.

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., & Lerchner, A. (2019). MONet: Unsupervised scene decomposition and representation. *International Conference on Learning Representations*.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chen, L., Xie, W., Liang, Y., He, H., Zhao, H., et al. (2026). Babyvision: Visual reasoning beyond language.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness [arXiv:1811.12231]. *International Conference on Learning Representations*.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hénaff, O. J., Koppula, S., Alayrac, J.-B., van den Oord, A., Vinyals, O., & Carreira, J. (2021). Efficient visual pretraining with contrastive detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The concept of object files: A tool for visual cognition. *Cognitive Psychology*, *24*(2), 175–219.

Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, *15*(4), 483–524.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, B., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Li, X., Yin, X., Li, C., Zhang, P., Zhang, X., Xiao, L., Zhang, J., & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. *Computer Vision – ECCV 2020*, *12375*, 121–137.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., & Kipf, T. (2020). Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*.

Mandler, J. M. (1992). How to build a baby: II. conceptual primitives. *Psychological Review*, *99*(4), 587–604. https://doi.org/10.1037/0033-295X.99.4.587

Maron, O., & Lozano-Pérez, T. (1998). A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Oikarinen, T., & Weng, T.-W. (2022). CLIP-Dissect: Automatic description of neuron representations in deep vision networks [ICLR 2023 Spotlight].

Orhan, A. E., Gupta, V. V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020 (neurips 2020)*. https://proceedings.neurips.cc/paper/2020/hash/7183145a2a3e0ce2b68cd3735186b1d5-Abstract.html

Orhan, A. E., & Lake, B. M. (2024). Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, *6*(3), 271–283.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Sermanet, P., Lynch, C., Hsu, J., & Levine, S. (2017). Time-contrastive networks: Self-supervised learning from video. *arXiv preprint arXiv:1704.06888*.

Smith, L. B. (2003). Learning to recognize objects. *Psychological Science*, *14*(3), 244–250.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). Saycam: A large, longitudinal audiovisual dataset recorded from an infant's perspective. *Open Mind*, *5*, 20–29. https://doi.org/10.1162/opmi_a_00039

Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5100–5111.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems*, *30*, 6306–6315.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3733–3742.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, J., Wang, L., Choi, Y., & Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10186–10196.

Zhao, J., Li, T., Jiang, D., Wu, S., Ramirez, A., & Lee, T. S. (2025). Perceptual inductive bias is what you need before contrastive learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9621–9630.